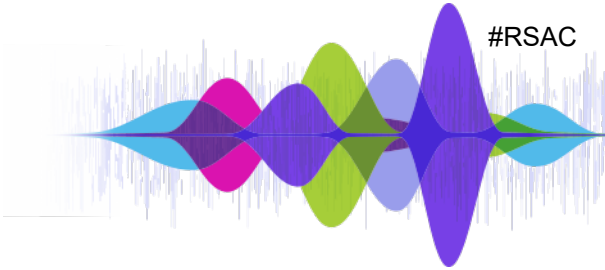
RSAC 2025 Conference

Many Voices.
One Community.

SESSION ID: FND-M02

Principles of GenAI Security: Foundations for Building Security In

Diana Kelley
CISO
Protect AI
https://www.linkedin.com/in/dianakelleysecuritycurve/



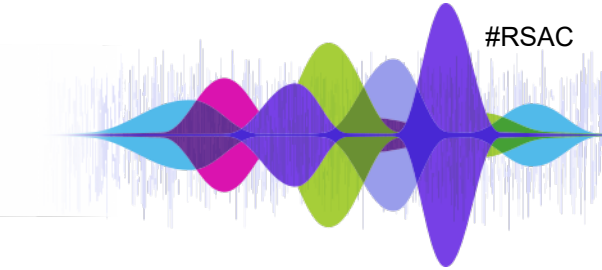
Disclaimer

This presentation is for educational purposes only and does not constitute legal, compliance, or security advice. Views expressed are those of the presenters and may not reflect Protect AI's positions. Attendees should conduct their own assessments before applying any information shared. Protect AI is not responsible for any actions taken based on this presentation.

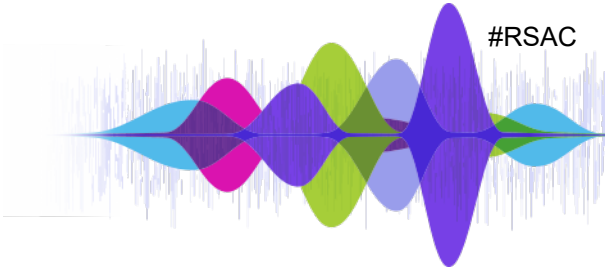
Attendees should note that sessions may be audio- or video-recorded and may be published in various media, including print, audio and video formats without further notice. The presentation template and any media capture are subject to copyright protection.

© 2025 RSA Conference LLC or its affiliates. The RSAC and RSAC CONFERENCE logos and other trademarks are proprietary. All rights reserved.

Agenda



- GenAI Explained
- GenAI Threat Landscape
- GenAI Security Architecture Considerations
- GenAI Security at Runtime
- Apply What you've Learned!

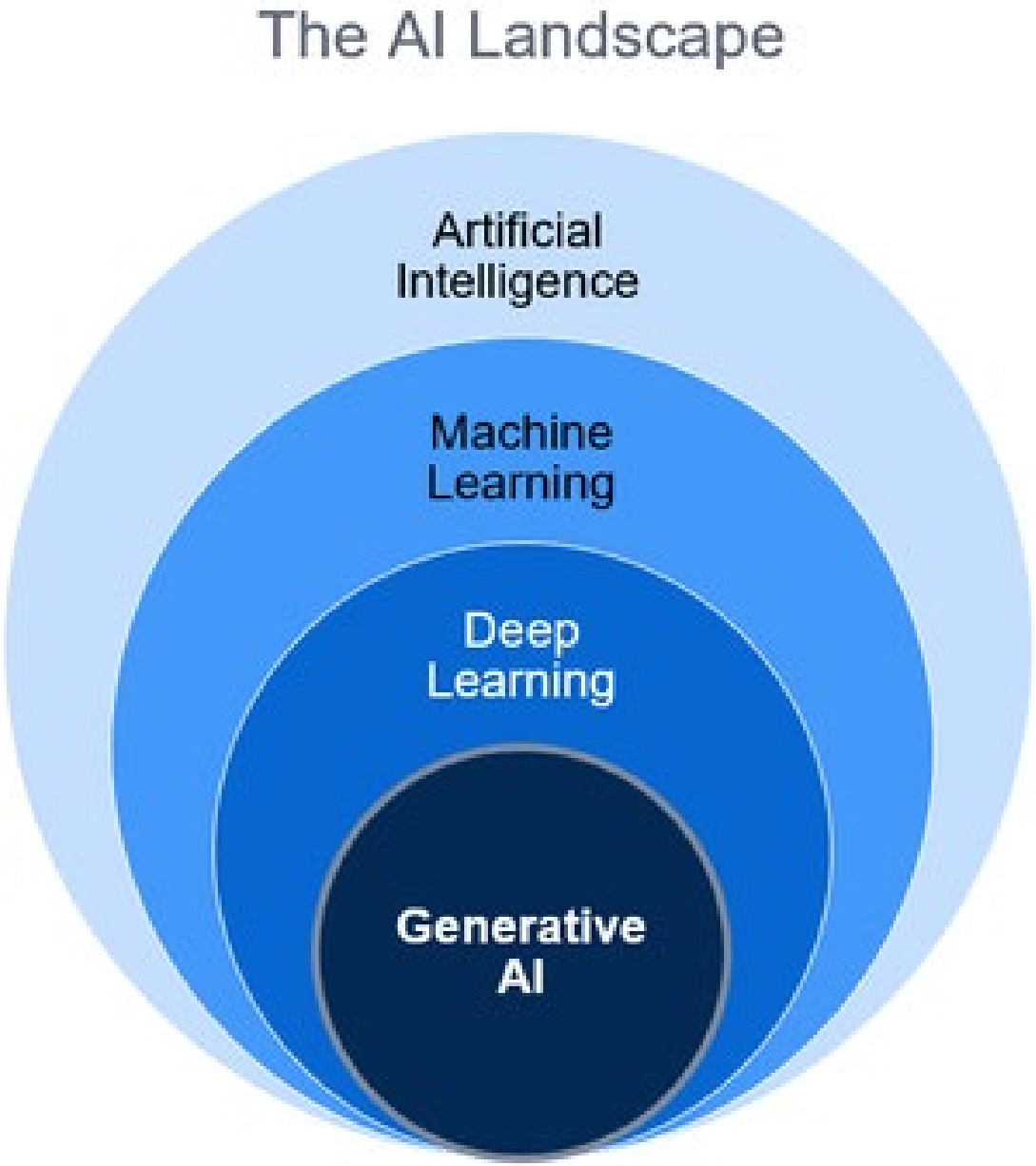


"*Discussions of artificial intelligence (AI) often swirl with mysticism regarding how an AI system functions. The reality is far more simple:*

AI is a type of software system."

CISA

Machine Learning powers Generative AI



Generative AI/GenAI

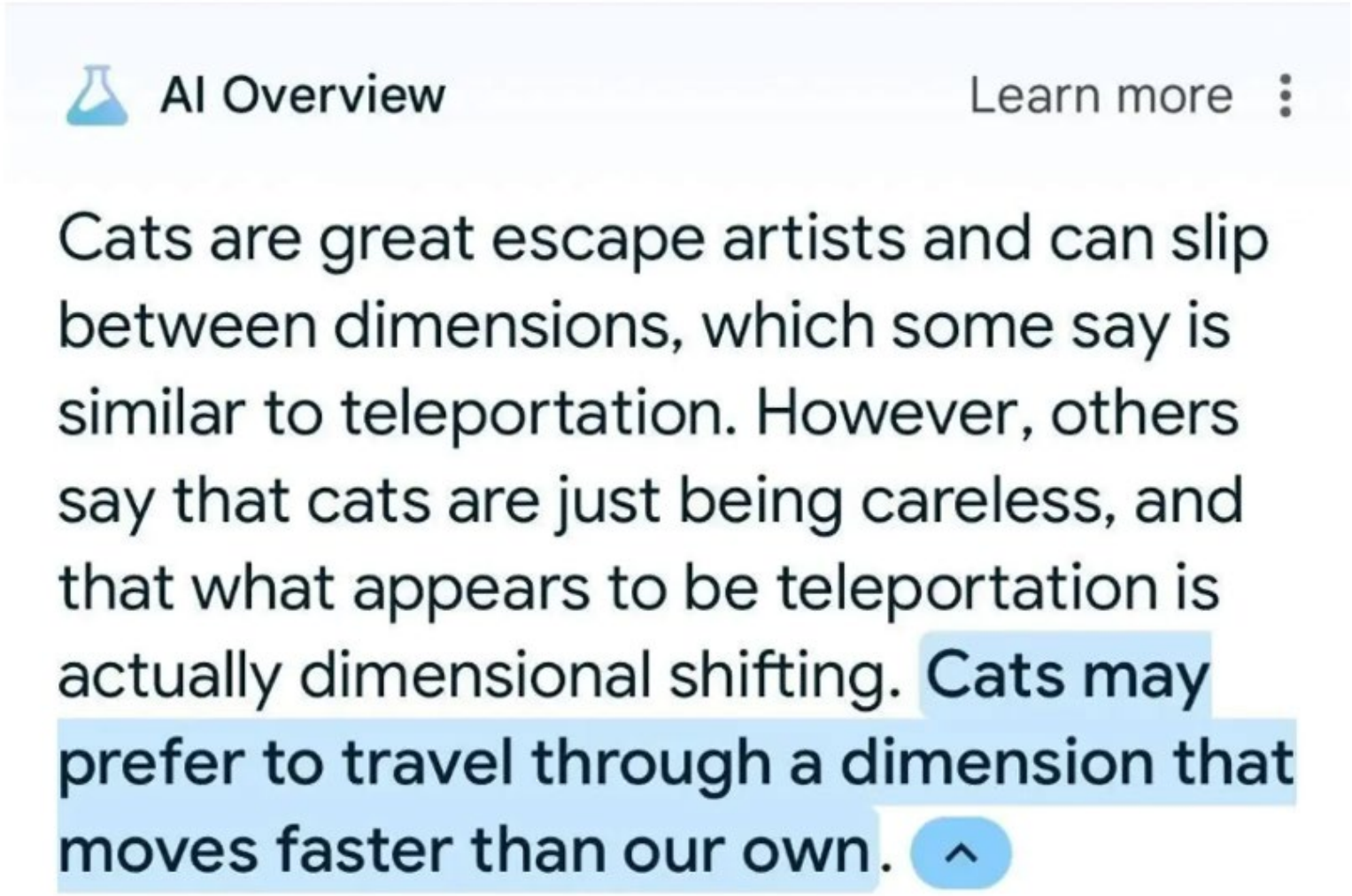
- What?
  - Trained on a corpus of knowledge
    - Ex: Large Language Models (LLM)
    - Corpus can be broad or specific
    - Fine Tuned or RAG (more on this in a bit!)
  - Generates new content based on that corpus
    - Statistical probability - Some call it a form of “autocorrect” or “type ahead” on steroids
    - Can be chained together for complex functions and “reasoning”
- Also; Images, Audio

GenAI Security Use Cases

- Summarization
  - Vendor Reviews: Populate SIGs from ingested docs (SOC2, Pen Tests)
  - Reporting
- Brain Storming/Extra “Eyes”
  - Threat Modeling
  - Policy Templates
- NLP Agents
  - Policy Bot
  - Interactive Security Awareness Training

Generative “Hallucinations”

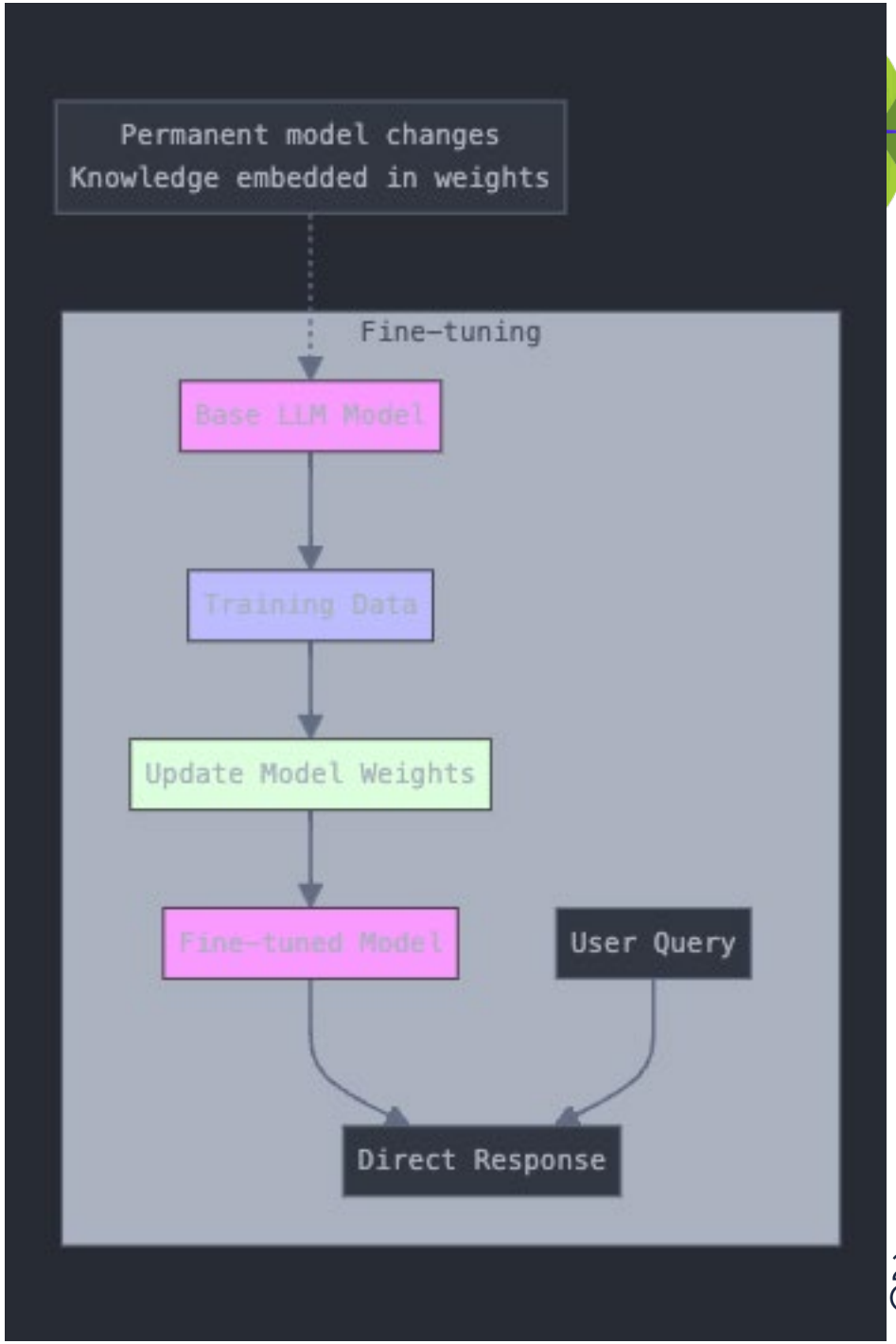
- Not the best for answering questions, if you can’t assess accuracy of the response…


AI Overview
Learn more
Cats are great escape artists and can slip between dimensions, which some say is similar to teleportation. However, others say that cats are just being careless, and that what appears to be teleportation is actually dimensional shifting. Cats may prefer to travel through a dimension that moves faster than our own.

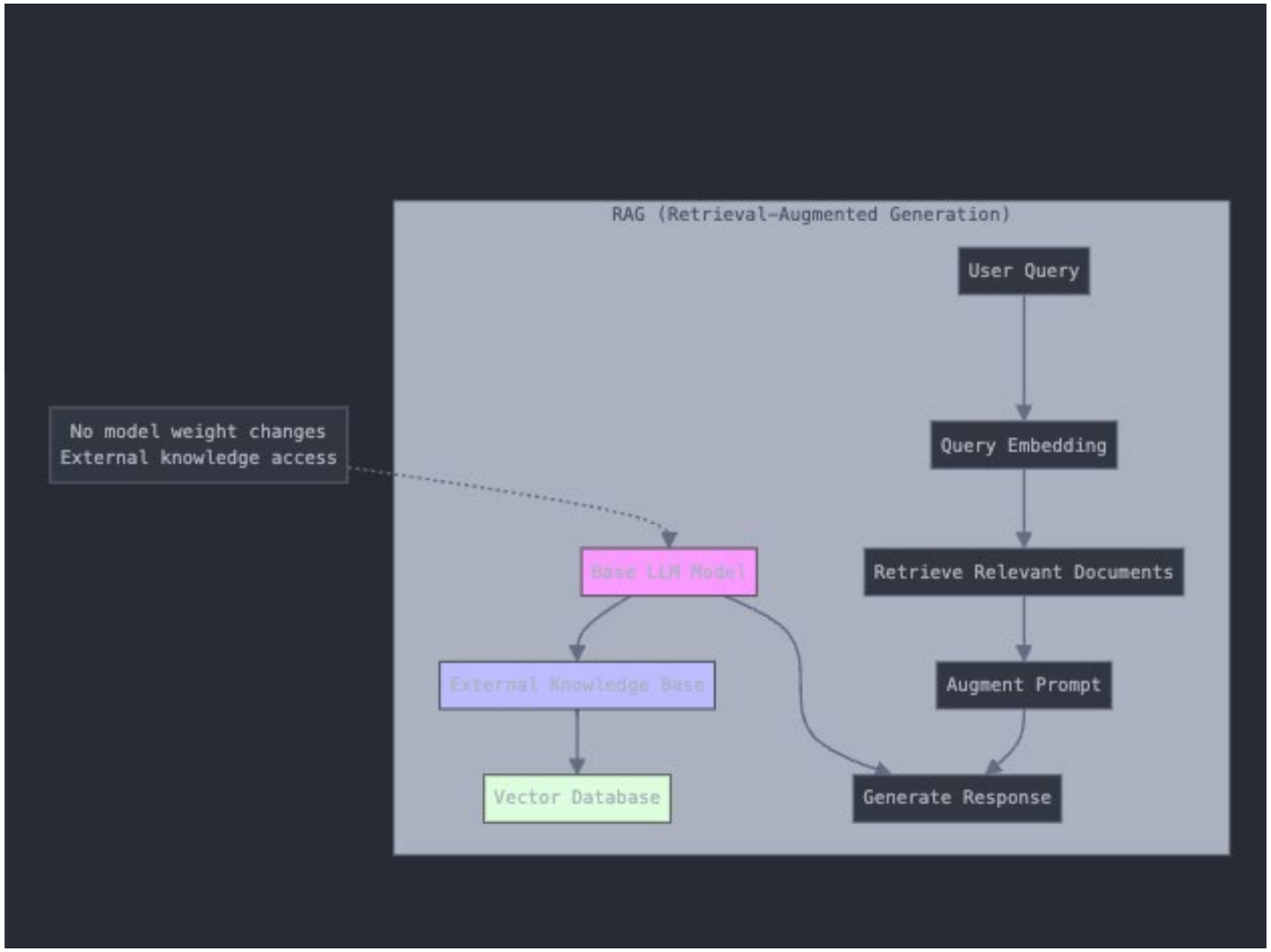
Query: How do cats teleport? Screenshot: Matt Novak/Google Search

Fine Tuning

- Modifies the base model's weights
- Incorporates knowledge directly into the model
- Faster inference since knowledge is built-in
- More efficient but less flexible to knowledge updates

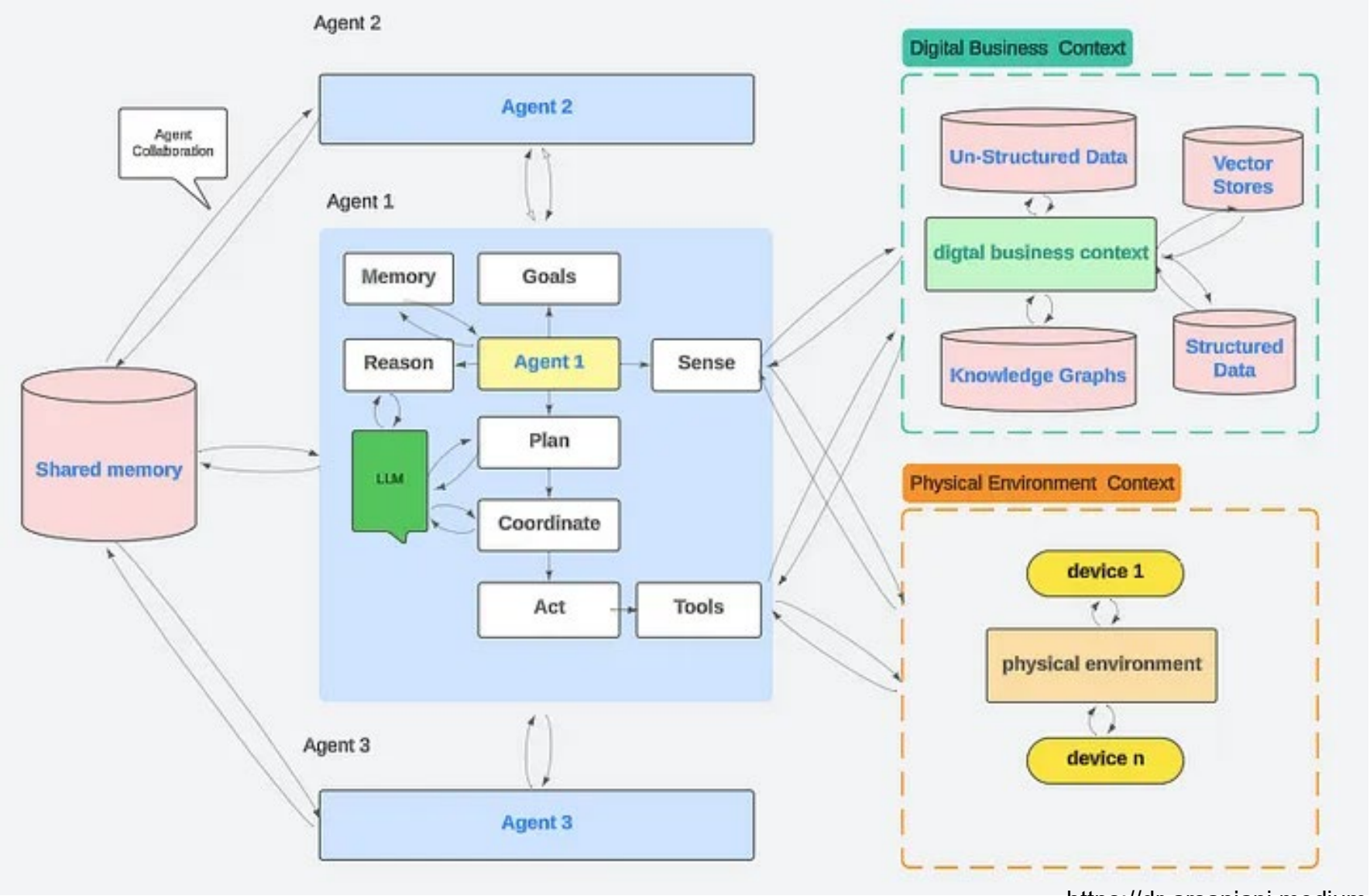


Retrieval Augmented Generation (RAG) Tuning



- Keeps the base model unchanged
- Uses an external knowledge base that can be easily updated
- Retrieves relevant information at runtime
- Augments the prompt with retrieved information
- More flexible but potentially slower

Anatomy of Agentic AI



- Reasoning and iterative planning
- Autonomous capabilities
- To solve multi-step problems

https://dr-arsanjani.medium.com/the-anatomy-of-agentic-ai-0ae7d243d13c

4 Steps of Agentic AI

1. **Perceive**: AI agents gather and process data from various sources

2. **Reason**: A large language model acts as the orchestrator, or reasoning engine, that understands tasks, generates solutions and coordinates specialized models for specific functions.

3. **Act**: By integrating with external tools and software via application programming interfaces, agentic AI can quickly execute tasks based on the plans it has formulated.

4. **Learn**: Agentic AI continuously improves through a feedback loop, or "data flywheel," where the data generated from its interactions is fed into the system to enhance models.

- Ex: Financial planner bot that helps decide payment plan of action for student loan repayment.

Original Source: https://blogs.nvidia.com/blog/what-is-agentic-ai/

RSAC 2025 Conference

GenAI Threat Landscape

What's Different?

Trust	Computer responses seen as more reliable or secure.
Impacts	AI and ML are being used for socially impactful purposes: medical, judicial, social.
Data Fuel	ML needs data to reason over, but the kind of data and how classification changes outcomes.
Visibility	AI and ML may be embedded but users don't know. CSR: human or bot?
Learning Curve	Lack of understanding and shared, common taxonomy across stakeholders.

Frameworks: Mitre and NIST

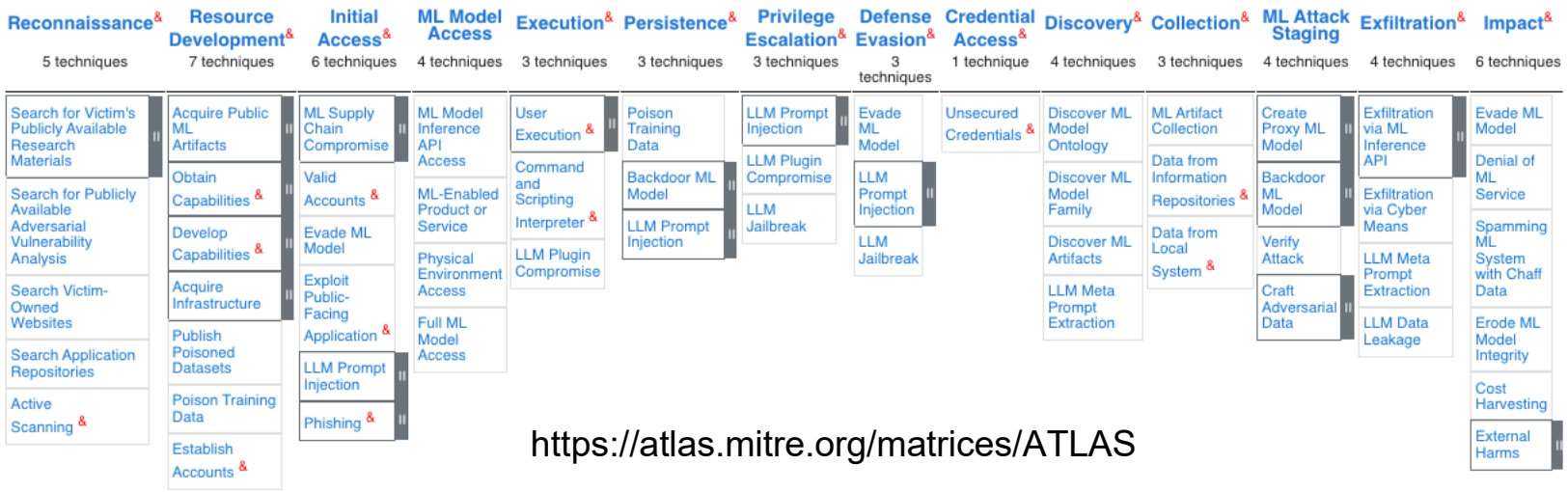
NIST

Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile

NIST AI 600-1

https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf

MITRE ATLAS

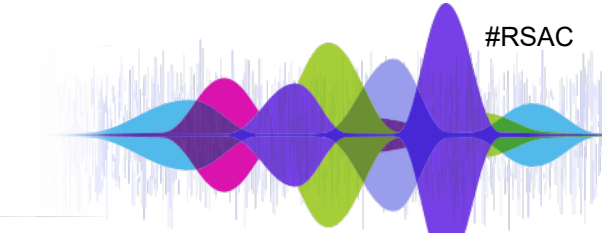


https://atlas.mitre.org/matrices/ATLAS

Top 10s and More

- LLMs
  - Top 10
  - Security and Governance Checklist
  - Security Verification Standard
- ML Top 10

https://genai.owasp.org/



LLM01: 2025
Prompt Injection

LLM01:2025 Prompt Injection

A Prompt Injection Vulnerability occurs when user prompts alter the...

Read More

LLM02: 2025
Sensitive Information Disclosure

LLM02:2025 Sensitive Information Disclosure

Sensitive information can affect both the LLM and its application...

Read More

LLM03: 2025
Supply Chain

LLM03:2025 Supply Chain

LLM supply chains are susceptible to various vulnerabilities, which can...

Read More

LLM04: 2025
Data and Model Poisoning

LLM04:2025 Data and Model Poisoning

Data poisoning occurs when pre-training, fine-tuning, or embedding data is...

Read More

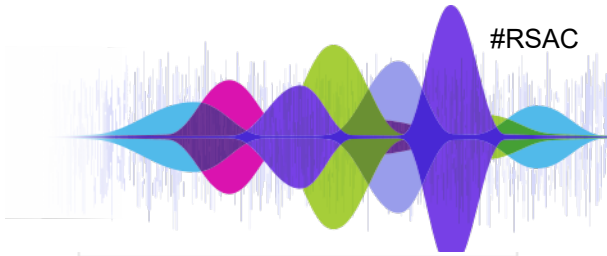
LLM05: 2025
Improper Output Handling

LLM05:2025 Improper Output Handling

Improper Output Handling refers specifically to insufficient validation, sanitization, and...

Read More

https://genai.owasp.org/llm-top-10/



LLM06: 2025
Excessive Agency

LLM06:2025 Excessive Agency

An LLM-based system is often granted a degree of agency...

Read More

LLM07: 2025
System Prompt Leakage

LLM07:2025 System Prompt Leakage

The system prompt leakage vulnerability in LLMs refers to the...

Read More

LLM08: 2025
Vector and Embedding Weaknesses

LLM08:2025 Vector and Embedding Weaknesses

Vectors and embeddings vulnerabilities present significant security risks in systems...

Read More

LLM09: 2025
Misinformation

LLM09:2025 Misinformation

Misinformation from LLMs poses a core vulnerability for applications relying...

Read More

LLM10: 2025
Unbounded Consumption

LLM10:2025 Unbounded Consumption

Unbounded Consumption refers to the process where a Large Language...

Read More

https://genai.owasp.org/llm-top-10/

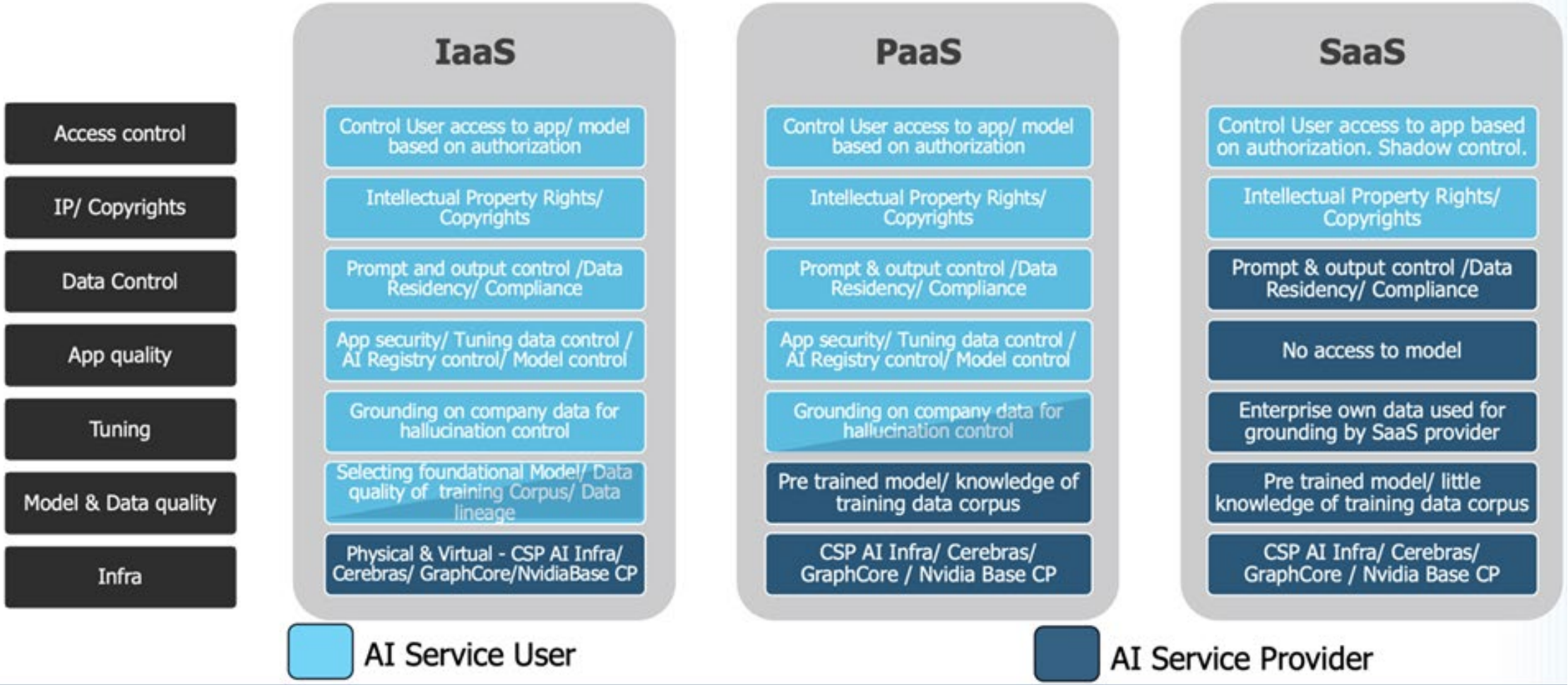
RSAC | 2025 Conference

GenAI Architectural Considerations

Input Vali

Many Voices.
One Community.

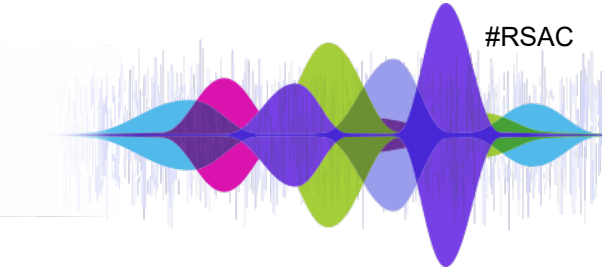
GEN-AI - SHARED RESPONSIBILITY MODEL


Access control
IP/ Copyrights
Data Control
App quality
Tuning
Model & Data quality
Infra
IaaS
Control User access to app/ model based on authorization
Intellectual Property Rights/ Copyrights
Prompt and output control /Data Residency/ Compliance
App security/ Tuning data control / AI Registry control/ Model control
Grounding on company data for hallucination control
Selecting foundational Model/ Data quality of training Corpus/ Data lineage
Physical & Virtual - CSP AI Infra/ Cerebras/ GraphCore/NvidiaBase CP
PaaS
Control User access to app/ model based on authorization
Intellectual Property Rights/ Copyrights
Prompt & output control /Data Residency/ Compliance
App security/ Tuning data control / AI Registry control/ Model control
Grounding on company data for hallucination control
Pre trained model/ knowledge of training data corpus
CSP AI Infra/ Cerebras/ GraphCore / Nvidia Base CP
SaaS
Control User access to app based on authorization. Shadow control.
Intellectual Property Rights/ Copyrights
Prompt & output control /Data Residency/ Compliance
No access to model
Enterprise own data used for grounding by SaaS provider
Pre trained model/ little knowledge of training data corpus
CSP AI Infra/ Cerebras/ GraphCore / Nvidia Base CP
AI Service User
AI Service Provider

Comparative shared responsibility

https://cloudsecurityalliance.org/blog/2023/07/28/generative-ai-proposed-shared-responsibility-model

GenAI & ZTA (Zero Trust Architecture)



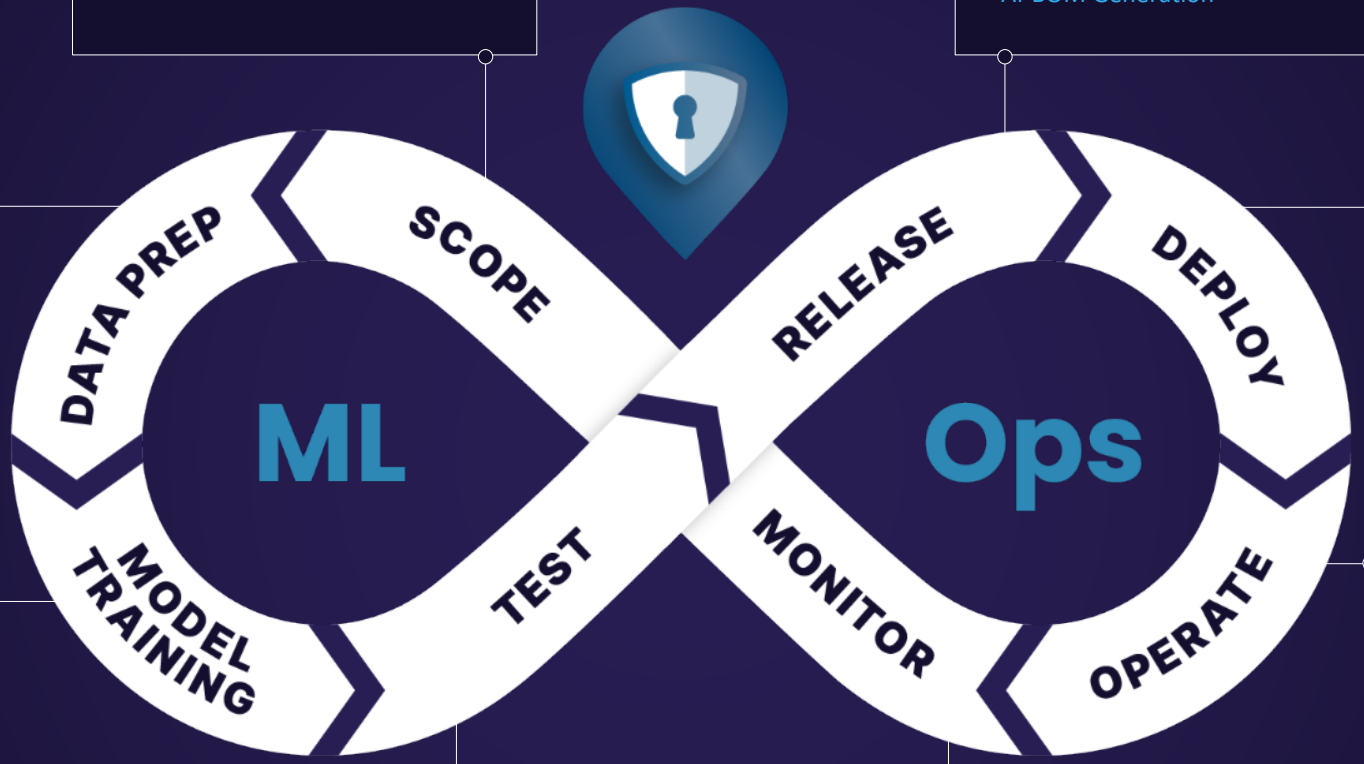
- AuthN & AuthZ
  - AI model interactions must be authenticated and authorized
  - treat each prompt or API call as potentially harmful
  - **Implement** with strong identity verification and access controls
- Data Handling & Sanitization
  - On inputs (prompts) and outputs (generated content)
  - **Implement** with validation layers to detect potential injection attacks and enforce allowed content policies (ex: types and formats)

GenAI & ZTA (Zero Trust Architecture)

- Segmentation
  - Between AI model components
  - **Implement** with microservices architectures and encrypted channels
- Monitoring & Logging
  - Continuously on all AI system behaviors, including prompt patterns and resource usage,
  - Establish baselines for normal operation
  - **Implement** automated detection of anomalies

GenAI & Secure by Design

- Design AI systems with the principle of least privilege
  - Each component should have only the minimum access and capabilities needed to perform its specific function.
  - Regularly review and rotate access credentials and implement strong key management practices.



- Task Identification
- Task Suitability
- Requirement Engineering
- Data Suitability
- Threat Modeling
- Policy Considerations

- Data Collection
- Baseline Establishment
- Data Formatting
- Data Cleaning
- Data Labelling
- Threat Modeling Refinement
- Privacy Requirements
- Secure Data Delivery

- Model Creation
- Model Training
- Error Analysis
- Performance Tracking
- Securing Notebooks
- Model Scanning
- Secure Model Gateway
- SCA & DAST

- Model Evaluation
- Testing Against Datasets
- Model Scanning
- LLM Benchmarking

- Model Acceptance
- Deployment Architecture (App/Container)
- Compliance Validation
- Digital Signing
- AI-BOM Generation

- Model Release to Production
- Policy Enforcement

- Provisioning
- Request Signing
- Rate Limiting
- Security Posture Refinement
- LLM Proxies
- LLM Security Scans

- Performance Evaluation
- Model Drift
- Input for Next Planning Phase
- Policy Refinements
- LLM Proxies
- LLM Audit and Governance

RSAC | 2025 Conference

GenAI Security at Runtime

Leverage the Architectural Work at Runtime

Monitoring for "Drift" and Accuracy

- Data Distribution
  - Real-world deviates from training data
  - Language use and cultural references evolve
  - Ex: A model trained before certain world events
- Behavioral Drift
  - Response consistency decreases over time
  - Ex: AI becomes more verbose than intended
- Performance Drift
  - Degradation in the quality or accuracy of generated content
  - Increased rate of hallucinations or factual errors
  - Ex: AI needs multiple retries to generate acceptable outputs

Monitoring for "Drift" and Accuracy

- Contextual Drift:
  - AI's understanding of context becomes less reliable
  - Decreased ability to maintain conversation coherence
  - Ex: AI might start losing track of conversation history more frequently
- Integration Drift:
  - Degradation in how the model interacts with external tools and APIs
  - Inconsistent handling of structured data
  - Ex: AI's API responses are less consistent in JSON formatting

GenAI Specific Forensic and Logging

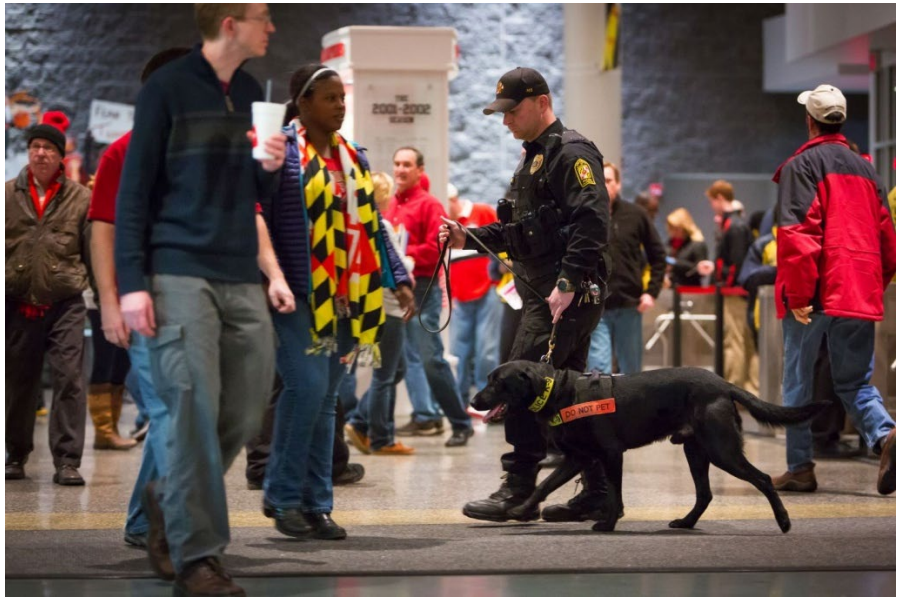
- Model Input-Output Correlation
  - Record and hash both input prompts and generated outputs to establish a clear audit trail
  - Things to capture
    - Unique transaction ID linking prompts to responses
    - Timestamp and user identity
    - Complete prompt history including system prompts
- Resource Consumption Patterns:
  - Track token usage per request
  - Monitor GPU/CPU utilization spikes
  - Log memory allocation and cache usage

GenAI Specific Forensic and Logging

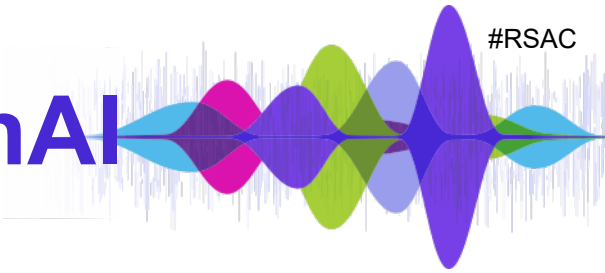
- Behavioral Analytics
  - Confidence scores and probability distributions
  - Hallucination indicators and model uncertainty
  - Number of retries and error rates
- Security Events and Boundaries
  - Log all prompt injection attempts
  - Track unauthorized access attempts
  - Monitor rate limiting triggers
- Data Provenance
  - Sources of training data updates
  - Document model version changes
  - Chain of custody for sensitive data

Incident Detection and Response for GenAI

- AI-Specific Detection Patterns
  - Monitor for prompt injection attempts and pattern anomalies
  - Detect unauthorized model access or unusual query patterns
  - Watch for data exfiltration through model outputs
  - Monitor for output degradation that could indicate poisoning



Incident Detection and Response for GenAI



- Response Procedures for AI Incidents:
  - Isolate affected model instances or endpoints
  - Preserve forensic data including prompt histories and outputs
  - Analyze impacted training data and model weights
  - Review access logs and authentication records
  - Document model behavior changes and performance impacts
  - Evaluate potential data exposure or privacy breaches

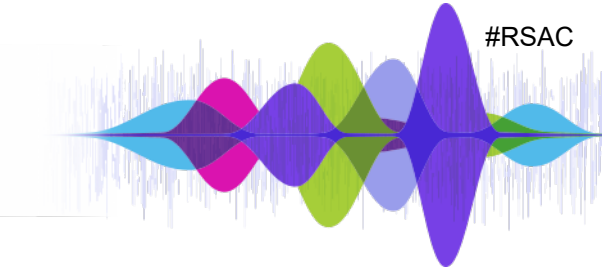
Incident Detection and Response for GenAI

- AI-Specific Incident Communication:
  - Notify affected users about potential data exposure
  - Document impact on model outputs and decisions
  - Report incidents to relevant regulatory bodies
  - Communicate with stakeholders about model reliability
  - Provide guidance on identifying compromised outputs
  - Share lessons learned with AI security community

GenAI AAR (After Action Reporting)

- Future Proofing w/ Prevention and Hardening
  - Implement enhanced prompt screening
  - Strengthen model access controls
  - Improve output validation and filtering
  - Add additional monitoring for similar attack patterns
  - Update incident response playbooks with new scenarios
  - Conduct AI-specific security training

Apply What You've Learned!



- Expand on Today's Learning!
  - NIST AI RMF - https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf
  - MITRE ATLAS - https://atlas.mitre.org/matrices/ATLAS
  - OWASP GenAI Security - https://genai.owasp.org/
- Get Hands On
  - Try testing GenAI resource to get started - https://gandalf.lakera.ai/baseline
  - Try installing and working with a foundational model locally - https://huggingface.co/docs/transformers/en/llm_tutorial
  - Complete a threat model for adoption of a public GenAI (SaaS) vs. self-hosted (P/IaaS)
- Go Deeper
  - The Developer's Playbook for Large Language Model Security: Building Secure AI Applications by Steve Wilson
  - Adversarial AI Attacks, Mitigations, and Defense Strategies: A cybersecurity professional's guide to AI attacks, threat modeling, and securing AI with MLSecOps by John Sotiropoulos
  - Generative AI Security: Theories and Practices by Ken Huang

RSAC | 2025 Conference

Many Voices.
One Community.

Thank you!!